UNIVERSITAT POLITÈCNICA DE CATALUNYA

MT-SUMMIT X, **September** *2005*

Bilingual N-gram Statistical Machine Translation

Universitat Politècnica de Catalunya

José B. Mariño, Rafael E. Banchs*, Josep M. Crego,
Adrià de Gispert, Patrik Lambert, José A. Rodríguez, Marta Ruiz

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Outline

1.- SMT and n-gram based SMT

2.- Tuple n-gram translation model

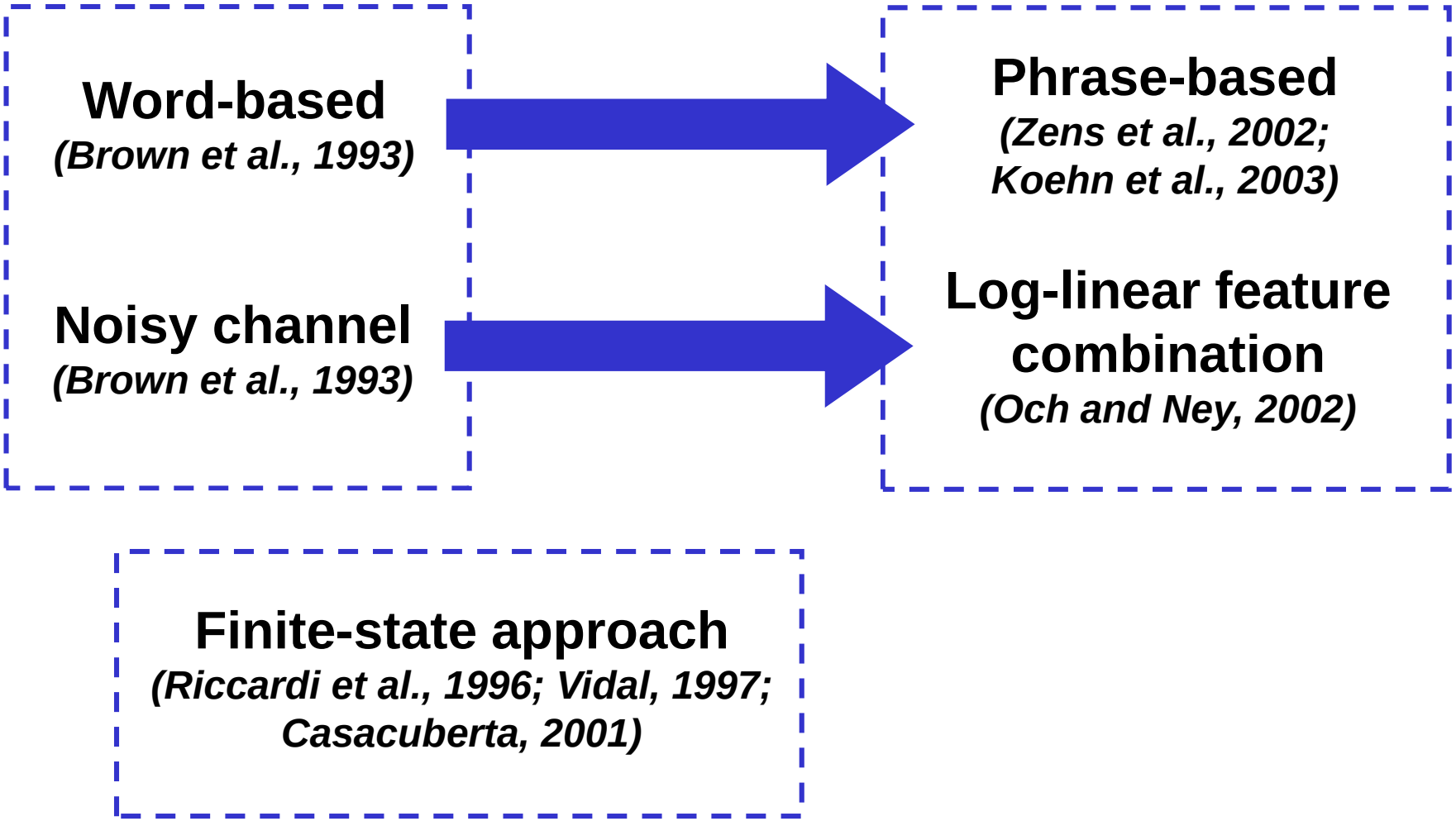
3.- Additional feature functions

4.- Overall SMT system description

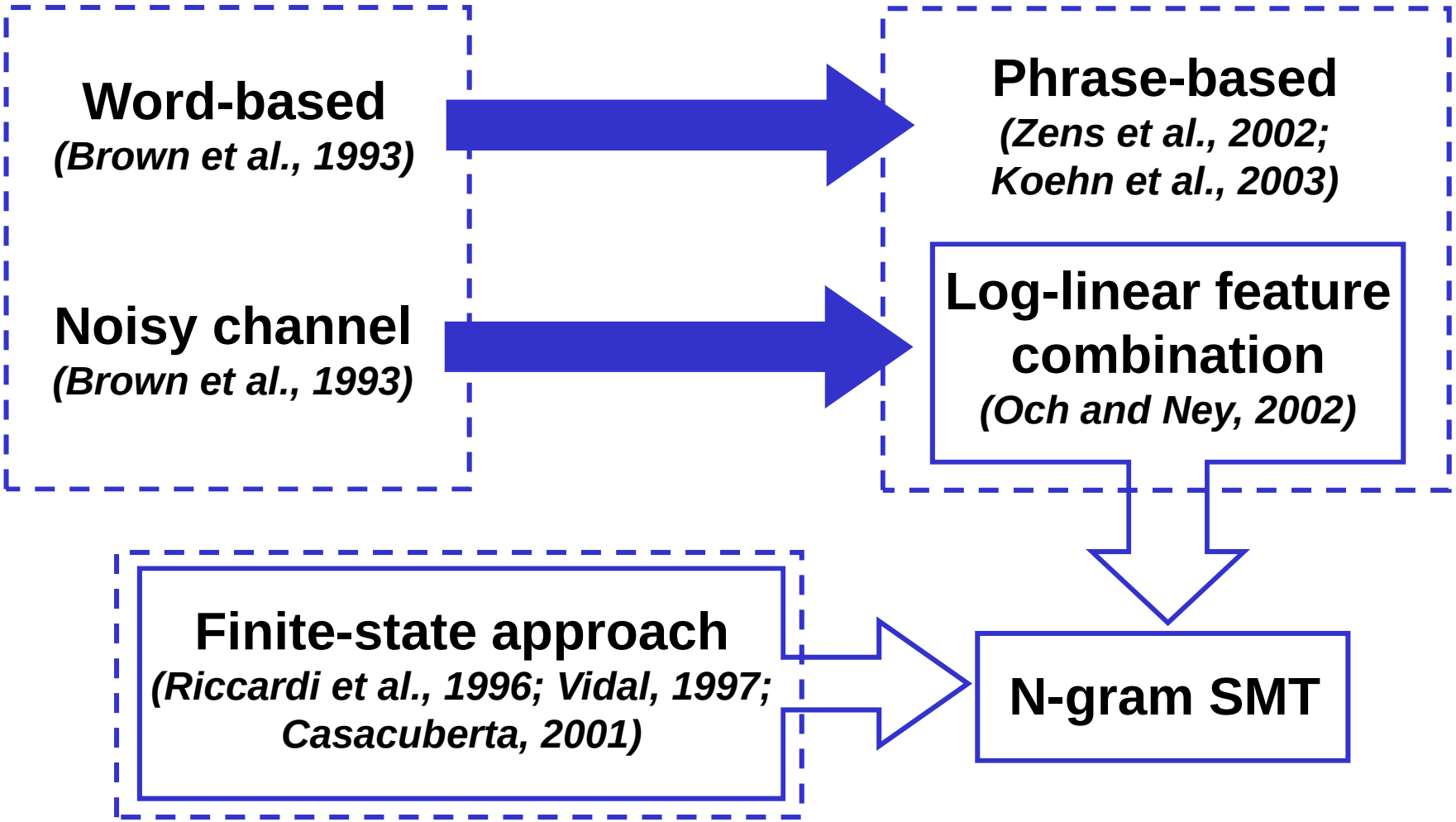
5.- EPPS translation experiments

6.- Conclusions and further work

Evolution of SMT



Origins of n-gram based SMT



The n-gram translation model

Implemented by using *3*-grams of bilingual units referred to as "tuples" ***(de Gispert et al., 2004)***

$$p_{TM}(T,S) \approx \prod_{k=1}^{K} p\left((t,s)_k \mid (t,s)_{k-1}, (t,s)_{k-2}\right)$$

- T → **target sentence**
- S → **source sentence**
- $(t,s)_k$ → **k^{th} tuple in the given sentence pair**

Tuple extraction criteria

Tuples are extracted according to the following criteria, which imply a unique segmentation, *(Crego et al., 2004):*

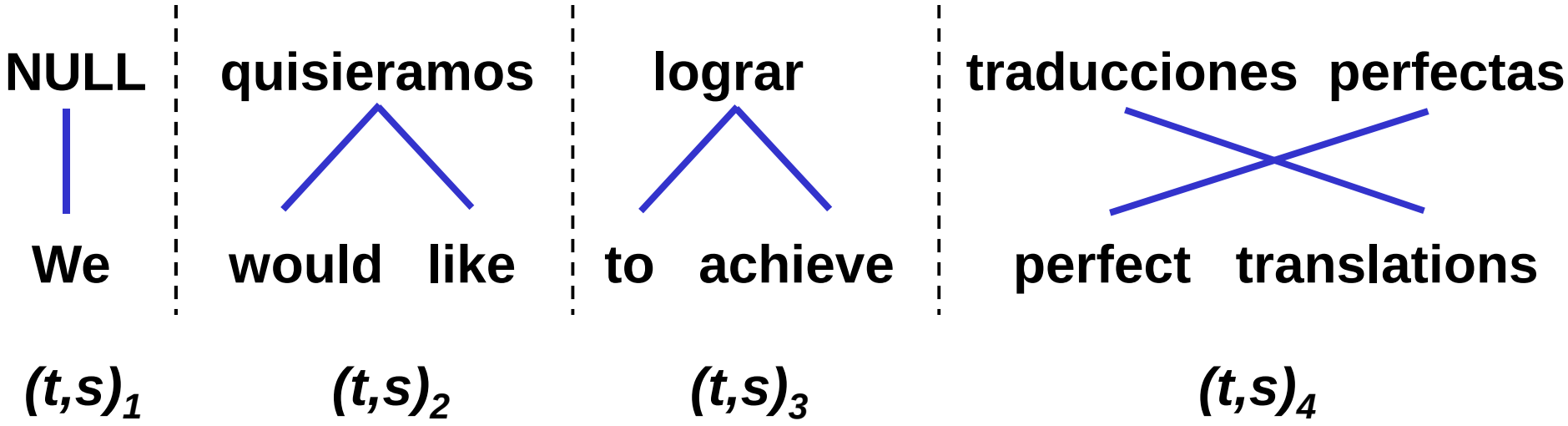
1.- No word inside a tuple is aligned to words outside the tuple

2.- A monotonic segmentation of bilingual sentence pairs should be produced

3.- Maximal segmentation, i.e. no smaller tuples can be extracted without violating the previous constraints

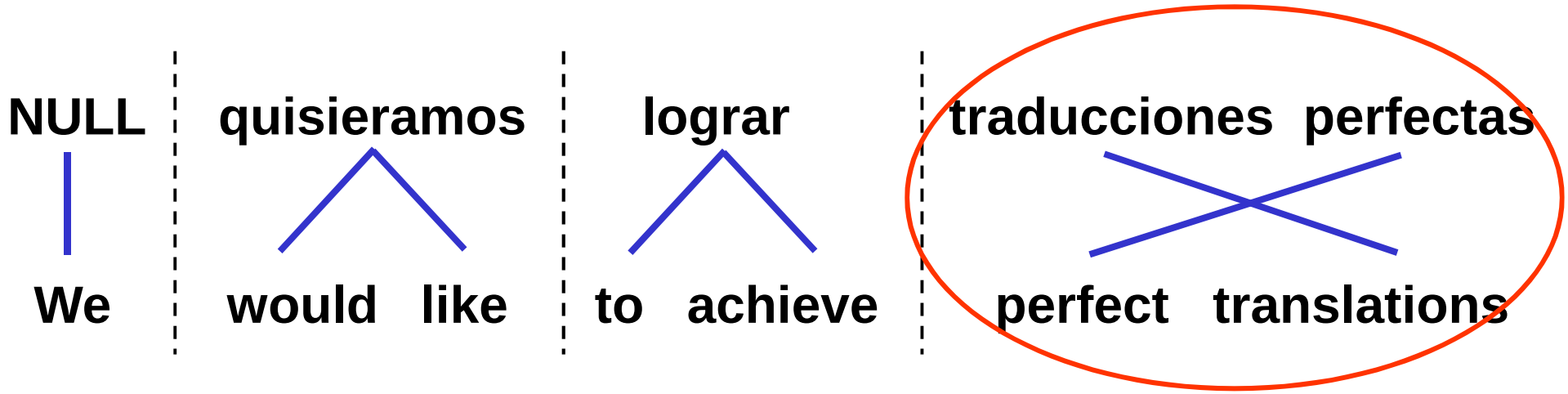
Example of tuple extraction

Tuples are extracted from the UNION set of source-to-target and target-to-source corpus alignments



Tuple n-gram model enhancements (1)

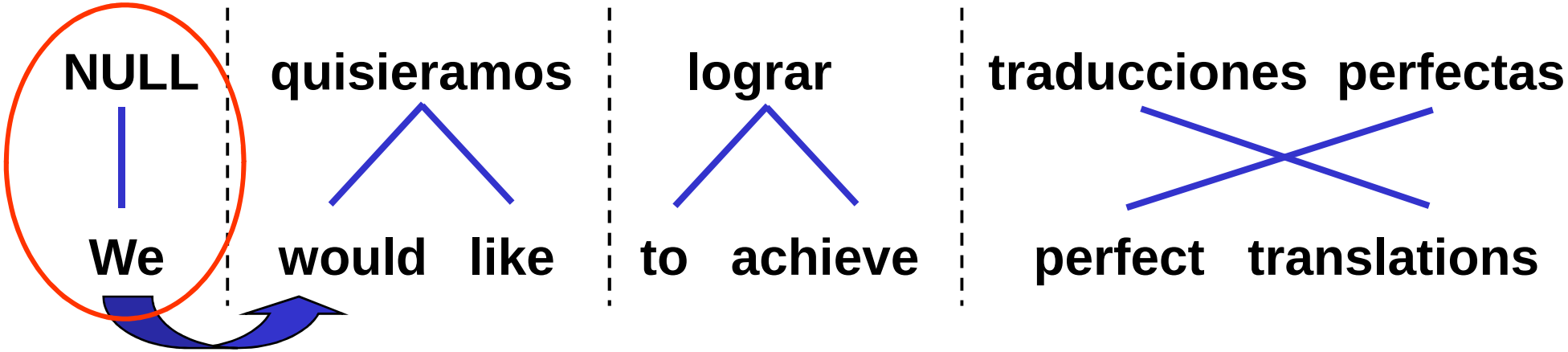
Embedded word dictionary



- ***1*-gram probabilities are added to the tuple n-gram model**
- **these probabilities are computed from the INTERSECTION set of source-to-target and target-to-source corpus alignments**

Tuple n-gram model enhancements (2)

NULL-to-word alignment cases



- **no NULL is expected to occur in a translation input**
- **target words that are linked to NULL are included into their corresponding following tuples**

Tuple n-gram model enhancements (3)

Tuple vocabulary pruning

- **allows for computational burden reduction while maintaining translation quality**
- **histogram pruning (only the *N* most frequent tuples with the same source sides are retained)**
- **the optimal pruning parameter *N* is adjusted empirically**

Target-oriented feature functions

- **Target language model (LM):** **a word *3*-gram**

$$p_{LM}(T_k) \approx \prod_{n=1}^{k} p(w_n \mid w_{n-1}, w_{n-2})$$

- **Word penalty model (WP):** **sentence length penalization**

$$wp(T_k) = \exp(\text{number of words in } T_k)$$

Lexicon-oriented feature functions

- **Forward lexicon model (FL): based on source-to-target IBM-1 lexical parameters *p(t|s)*** ***(Och et al., 2004)***

- **Backward lexicon model (BL): based on target-to-source IBM-1 lexical parameters *p(t|s)***

$$p_{XL}((t,s)_n) = (I+1)^{-J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(t_n^i \mid s_n^j)$$

Log-linear combination of models

$$\hat{T} \approx \underset{T}{argmax} \left\{ p_{TM}^{\lambda_{TM}} \quad p_{LM}^{\lambda_{LM}} \quad wp^{\lambda_{WP}} \quad p_{FL}^{\lambda_{FL}} \quad p_{BL}^{\lambda_{BL}} \right\}$$

- **λ_i's are empirically adjusted via an optimization procedure**
- **translation BLEU *(Papineni et al., 2002)* is maximized over a development data set**
- **optimization is based on simplex algorithm *(Press et al., 2002)***

UNIVERSITAT POLITÈCNICA DE CATALUNYA

N-gram based decoding

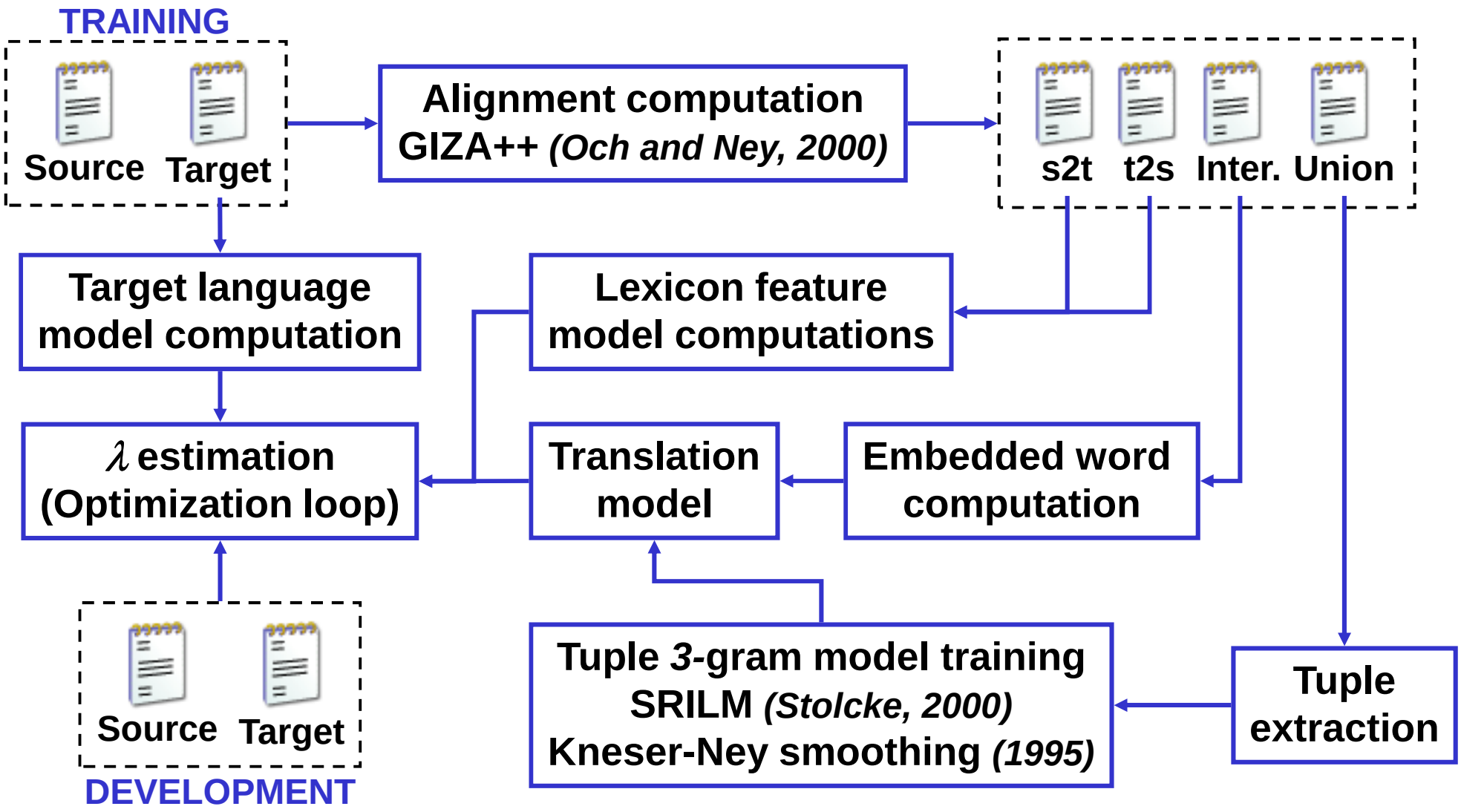
Specific n-gram based search engine ***(Crego et al., 2005)***

- **dynamic programming beam-search strategy**
- **considers the five described models simultaneously**
- **allows for threshold pruning, histogram pruning and hypothesis recombination**
- **monotonic search modality is used***

*** Recent results on n-gram decoding reordering capabilities in this afternoon presentation by J. M. Crego**

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Overall process description



Official Spanish and English transcriptions of the EPPS, available through the TC-STAR consortium

Set	Lang.	Sent.	Words	Vocab.	Av.SL	Ref.
Train	EN	1.22 M	33.4 M	105 K	23.7	1
	ES	1.22 M	34.8 M	169 K	28.4	1
Dev.	EN	1008	26.0 K	3.2 K	25.8	3
	ES	1008	25.7 K	3.9 K	25.5	3
Test	EN	1094	26.8 K	3.9 K	24.5	2
	ES	840	22.7 K	4.0 K	27.0	2

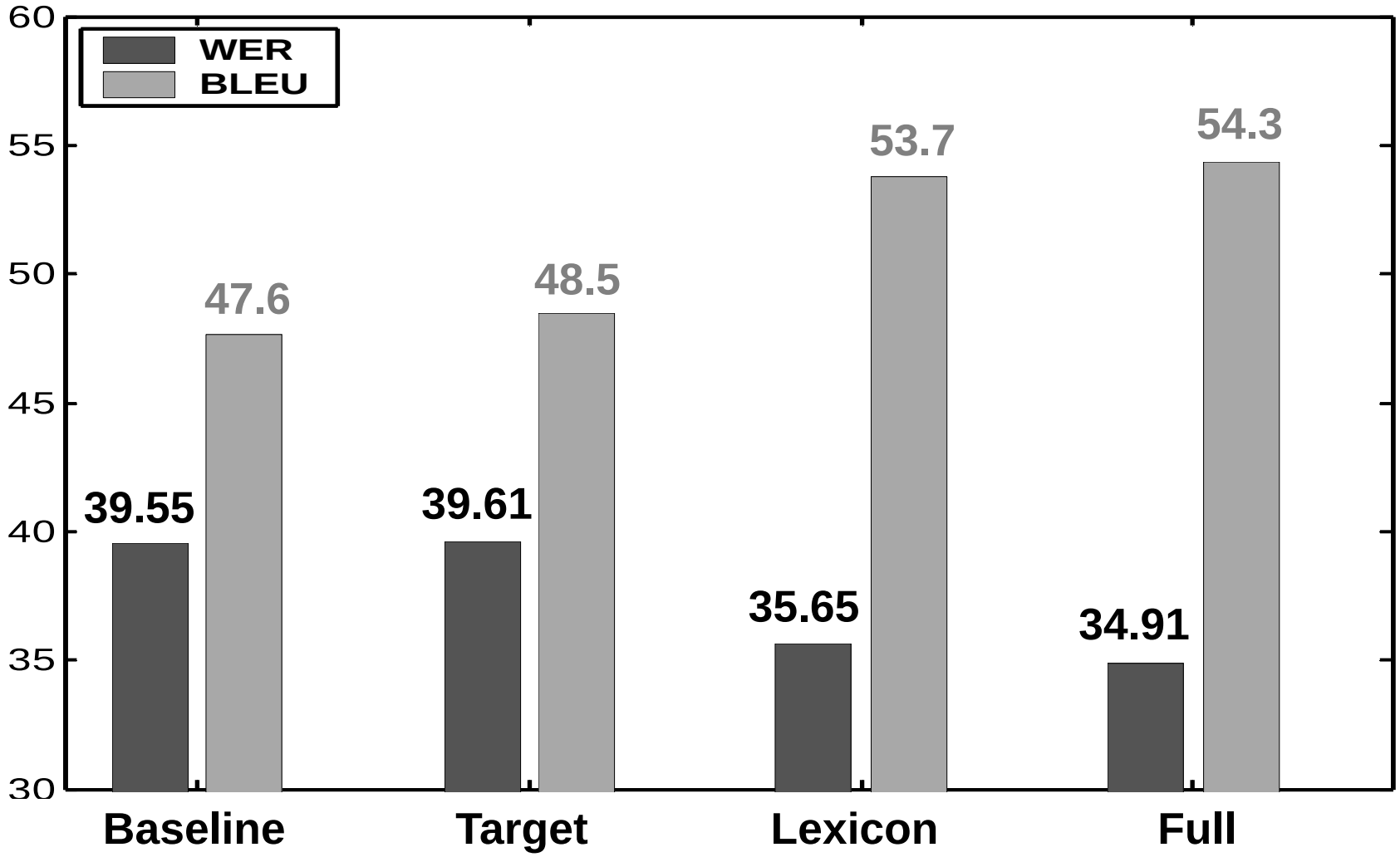
UNIVERSITAT POLITÈCNICA DE CATALUNYA

System configurations evaluated

	TM	LM	WP	FL	BL
Baseline system	✓				
Target-reinforced system	✓	✓	✓		
Lexicon-reinforced system	✓			✓	✓
Full system	✓	✓	✓	✓	✓

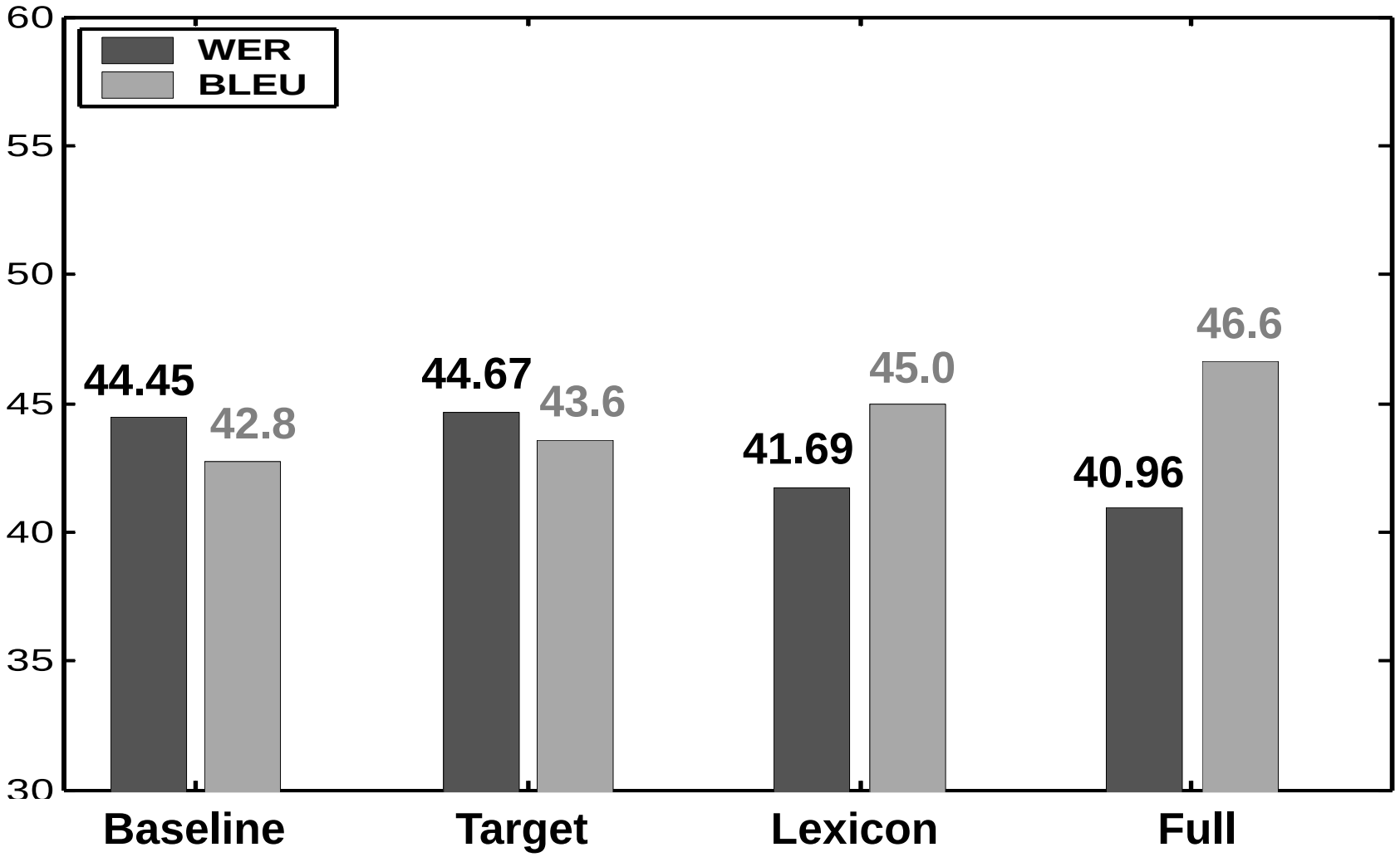
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Spanish-to-English translation results



UNIVERSITAT POLITÈCNICA DE CATALUNYA

English-to-Spanish translation results



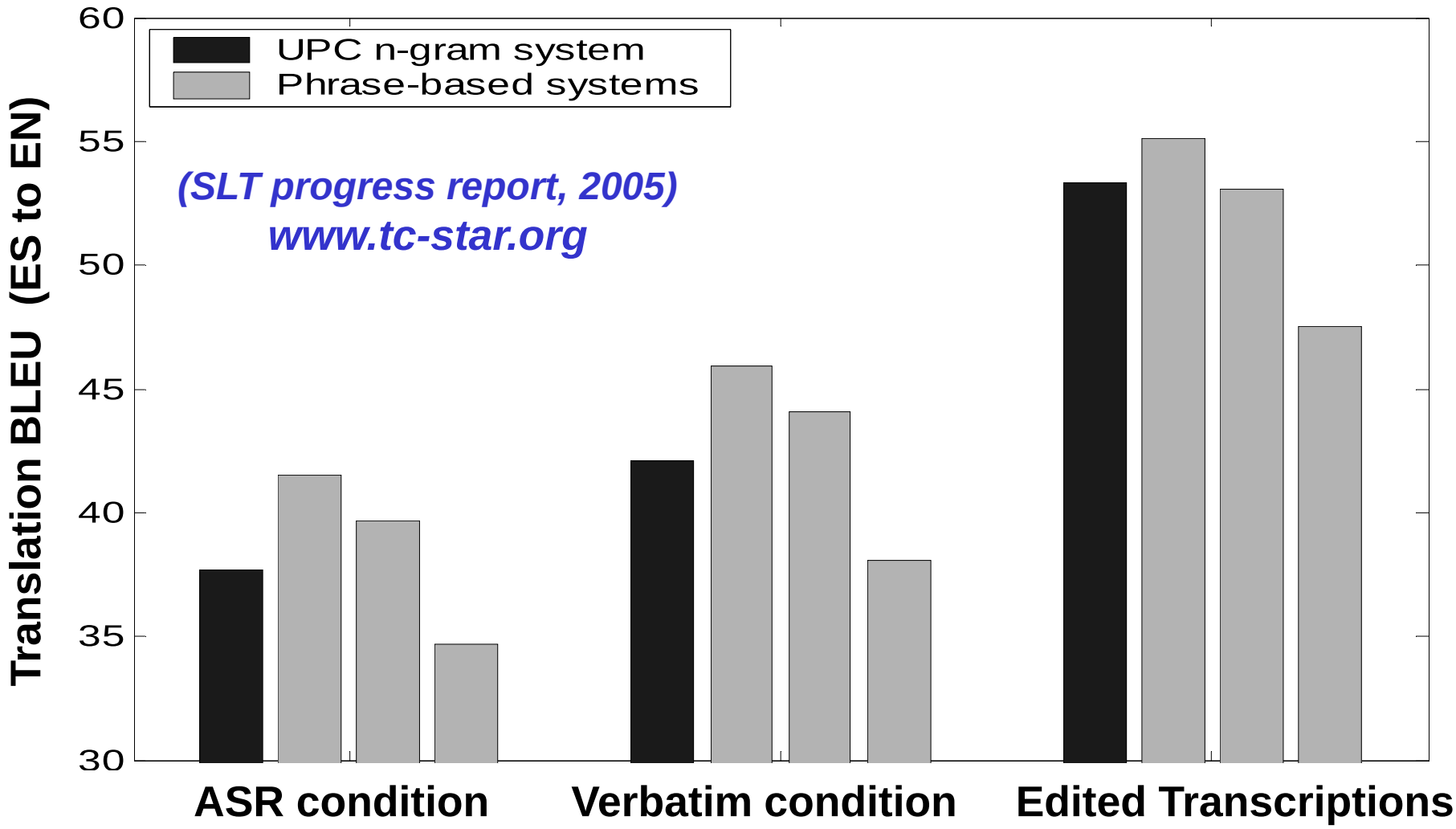
Translation error analysis

The most common errors encountered:

- **Verbal forms: wrong verbal tenses and auxiliary forms, active/passive voice related errors**
- **Omitted translations: large amount NULLs occurring**
- **Reordering problems: mainly on adjective-noun and subject-verb constructions**
- **Concordance problems: inconsistencies regarding gender and number**

UNIVERSITAT POLITÈCNICA DE CATALUNYA

N-gram SMT compared to phrase SMT



UNIVERSITAT POLITÈCNICA DE CATALUNYA

Conclusions

- **The tuple n-gram translation model, when used along with the four additional features presented, provides state of the art translations**

- **The inclusion of the additional feature functions produces an important improvement over the baseline system, being the lexicon models the ones with the most impact**

- **Spanish-to-English translations are consistently and significantly better than English-to-Spanish translations**

Further work

- **Linguistic information must be used to cope with the most common types of observed errors. In this sense, ongoing research at UPC is focused on using:**
  - **1.- bilingual word and multi-word dictionaries**
  - **2.- verbal form analysis and classification**
  - **3.- POS- tag and chunk information**

- **Reordering strategies, as well as non-monotonic decoding schemes, for the proposed SMT system must be developed***

*** Recent results on n-gram decoding reordering capabilities in this afternoon presentation by J. M. Crego**

UNIVERSITAT POLITÈCNICA DE CATALUNYA

MT-SUMMIT X, **September** *2005*

Bilingual N-gram Statistical Machine Translation

Universitat Politècnica de Catalunya

José B. Mariño, Rafael E. Banchs*, Josep M. Crego,
Adrià de Gispert, Patrik Lambert, José A. Rodríguez, Marta Ruiz

UNIVERSITAT POLITÈCNICA DE CATALUNYA

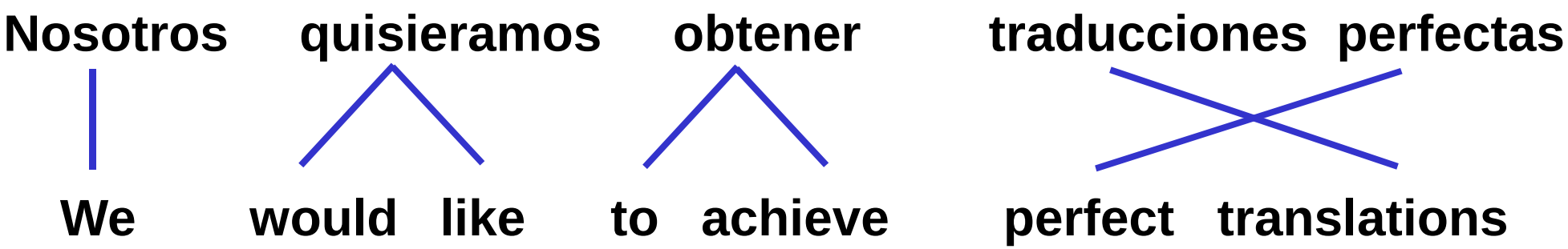
Relationship between tuples and phrases

Tuples actually constitute a subset of Phrases, the main differences between both SMT approaches are:

	TUPLES	PHRASES
Translation model probabilities	n-gram probabilities	relative frequency and other features
Resulting segmentation	unique and monotonic	non-unique and not necessarily monotonic
Bilingual unit length	depends on data alignments	maximum length is a parameter

Relationship between tuples and phrases

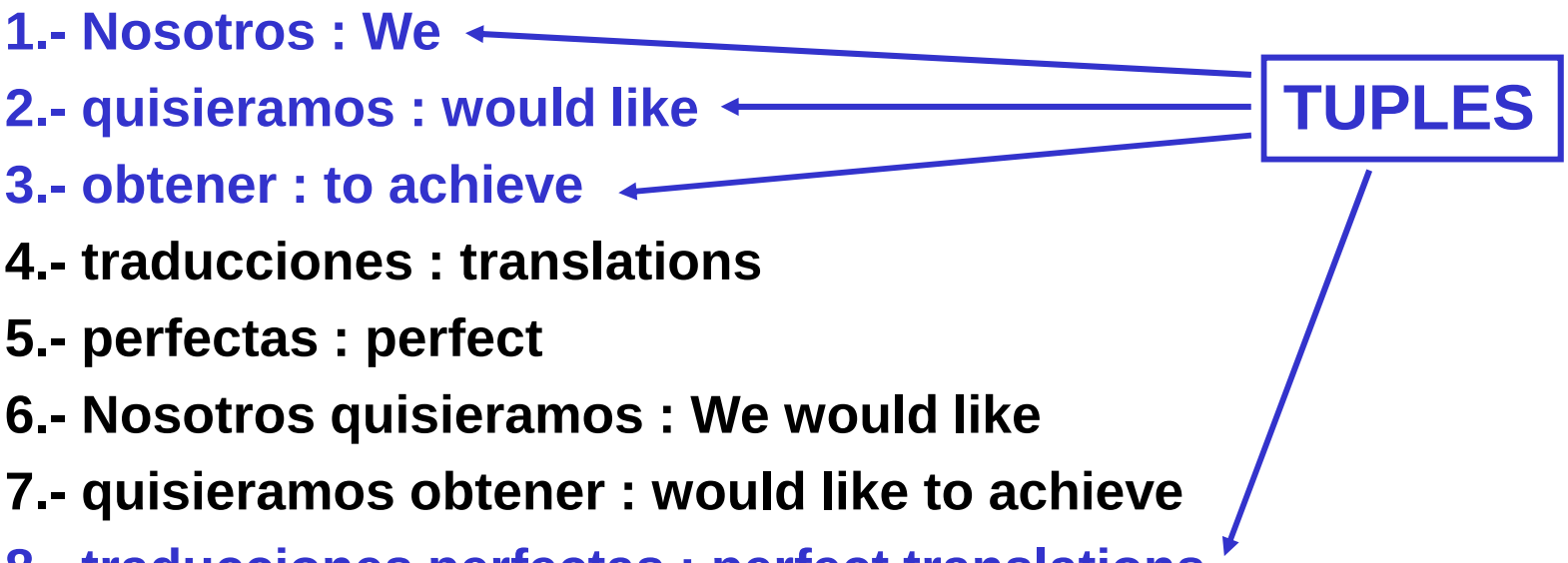
Consider the following example:



- **it contains only four tuples**
- **it contains a total of 12 phrases**

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Relationship between tuples and phrases



1.- Nosotros : We
2.- quisieramos : would like
3.- obtener : to achieve
4.- traducciones : translations
5.- perfectas : perfect
6.- Nosotros quisieramos : We would like
7.- quisieramos obtener : would like to achieve
8.- traducciones perfectas : perfect translations
9.- Nosotros quisieramos obtener : We would like to achieve
10.- obtener traducciones perfectas : to achieve perfect translations
11.- quisieramos obtener traducciones perfectas : would like to achieve perfect translations
12.- Nosotros quisieramos obtener traducciones perfectas : We would like to achieve perfect translations

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Total number of tuple n-grams

Direction	*1*-grams	*2*-grams	*3*-grams
ES to EN	2.040 M	6.009 M	1.798 M
EN to ES	2.023 M	6.092 M	1.747 M

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Optimized λ values

Direct.	System	λ_{TM}	λ_{LM}	λ_{WP}	λ_{FL}	λ_{BL}
ES to EN	Baseline	1.00				
	Target	1.00	0.30	0.32		
	Lexicon	1.00			0.48	0.06
	Full	1.00	0.48	0.28	0.48	0.13
EN to ES	Baseline	1.00				
	Target	1.00	0.32	0.26		
	Lexicon	1.00			0.17	0.07
	Full	1.00	0.80	0.75	0.23	0.18

Translation samples

- **The policy of the European Union on Cuba NULL must [must not] change .**
- **To achieve these purposes , it is necessary NULL for the governments to be allocated [to allocate] , at least , 60 000 million NULL dollars a year ...**
- **In the UK we have NULL [already] laws enough [enough laws] , but we want to encourage NULL other States ...**